Alpes4science project : SMS corpus processing and tokenization problems

Eleni Kogkitsidou and **Georges Antoniadis**
Lidilem Laboratory, University of Stendhal,
Grenoble, France

{eleni.kogkitsidou,georges.antoniadis}@u-grenoble3.fr

Abstract

Virtual textual communication involves numeric supports as transporter and mediator. SMS language is part of this type of communication and represents some specific particularities. An SMS text is characterized by an unpredictable use of white-spaces, special characters and a lack of any writing standards, when at the same time stays close to the orality. This paper aims to expose the database of alpes4science project from the collation to the processing of the SMS corpus. Then we present some of the most common SMS tokenization problems and works related to SMS normalization.

1 Introduction

With the appearance of new forms of virtual communication (chats, email, social networks, etc.), new terms have been invented to describe this new type of communication: computer-mediated communication (CMC), written computer-mediated communication or network-mediated communication, cybercommunication, netspeak, etc. Since 90s, SMS communication belongs to this type of communication and it's the subject of our study. The interest to study the SMS communication and the SMS language, in our case, is identified at the particularities which this language presents. It's a discourse that escapes the institutional constraints and lacks any standards (Panckhurst, 2009). As it is mentioned by Barasa and Mous (2009), SMS text is characterized by a rich lexical creativity without conventions, and a creation of a new form of orthography. Stark (2011) described SMS as a strict and particular writing code which combines several methods to shorten sentences and words. On the other side, it is close to the orality by remaining a written form and that's why this kind of language is a subject of interest for many researchers (Antoniadis et al., 2011).

2 The alpes4science project

The observation of these particularities requires authentic and certified materials in order to obtain an objective view (Fairon and Paumier, 2006). The sms4science[1] project aims to respond to this need by launching, in 2004, the first collation of SMS at CENTAL[2] laboratory of Catholic University of Louvain, and establishing a collation methodology and protocols for SMS corpora construction. Since then, several other works related to this project have been released (Reunion Island, 2008, http://www.lareunion4science.org/; Switzerland, 2009, http://www.sms4science.ch/; Quebec, 2010, http://www.texto4science.ca/; Montpellier, 2010, http://www.sud4science.org/) (Panckhurst, 2013).

Our study uses as starting point the SMS corpus of alpes4science[3] project which is the part of sms4science project. The alpes4science project was signed in 2009 between LIDILEM[4] and the

[1] http://www.sms4science.org/
[2] Center of Natural Language Processing
[3] www.alpes4science.org
[4] lidilem.u-grenoble3.fr/

General Council of Hautes-Alpes for the purpose to create a database.

The collation took place from 1 October 2010 to 31 January 2011 in Hautes-Alpes and Isère of France. For this reason, the topic of messages is related to local and seasonal events (snow, ski, pistes, end of year celebrations, greetings etc.). However, we identify some sent messages which were saved in the mobile phone and they are not related to the chronological period of the collation, such as for example messages like : "thanks", "see you later" (Chabert et al., 2012).

In total, 359 people sent their 22054 SMS to the platform. Each participant should send his messages to a special number by writing the "SMS05" code at the beginning of every sent message. Thereafter, all messages were transported to a special dedicated platform. The registration was done once the participant had sent his first message beginning with the "SMS05" code and following his phone number. In this way, participants were automatically associated with an identification number and they could transfer their messages (Antoniadis et al., 2011).

The participants of the project were invited to complete a questionnaire with varied information concerning their social profile (age, gender, education level, profession, mother tongue etc.), as well as, their communicative character (texting frequency, keyboard, language register etc.). Among participants 119 persons didn't answer the questionnaire. As for the rest of 240 persons we know that the 70.8% represents female SMS writers and the 29.2% male writers aged from 14 to 69 years old. This metadata is an incontestable material for the production of scientific studies through the analysis of this information in the fields of linguistics, natural language processing, sociology and sociolinguistics for the purpose of establishing actual observations.

2.1 Corpus Processing

With the construction of the SMS corpus we can examine adequately the function of languages and explore exhaustively authentic language productions. In our case, we focus on the original SMS corpus which allows us to examine the particularities of this type of communication. There are two types of treatment that are essential to make the SMS corpus operational and able to give way to other NLP applications (Sproat et al., 2001; Beaufort et al., 2010) : the *anonymization* of sensitive data for ethical reasons and the *transcription* that aims to make readable and usable messages in order to facilitate the operation of the corpus.

2.1.1 Corpus anonymization

The anonymization of data doesn't exclusively concern SMS messages but also any other form of communication and data type (state protected data, University restricted or critical data, telecommunications, electronic commerce, etc.). This is a compulsory process by ethics and by agreement with the CNIL[5] (1442138) for the authorized diffusion of corpora in order to preserve the confidentiality of transmitted information. In alpes4science corpus we consider as sensitive data: last names, nicknames, surnames, phone numbers, e-mail addresses, URL, codes, postal addresses, as well as, any other information which allows the indirect identification a person. The anonymization process had been achieved via a web interface designed for this project which was capable to detect standard format data (for example: e-mail addresses, URL, phone numbers), then, three researcherswere in charge to verify the result which were automatically produced. The data to be anonymized was replaced by a new form. This new form matched ***(DATA NAME)_Number of data character*** (table 1).

Original SMS	j'écris à Mathieu
Anonymized	j'écris à ***SURNOM_7***
Translation	I'm writing to Mathieu

Table 1: Anonymization example

2.1.2 Corpus transcription

The transcription of SMS aims to make a message which contains abbreviations, phonetizations, extensions etc. understandable to everyone. Before proceed to the SMS transcription we had defined, in a strictly way, through a protocol all the elements which meant to be modified from the original message to the standard language. The

[5]http://www.cnil.fr/english/

purpose of this processing is to release a minimum of changes and only if it is necessary (table 2).

Original SMS	Oui bien sur qan tu veu
Transcription	Oui bien sûr quand tu veux

Table 2: transcription example

The applied methodology consists of trancoding manually SMS which from their part contribute to create a dictionary to the database with SMS words. This method proposes subsequently to the researcher the possibility to make a choice to keep or change the word to by transcription via a web interface.

3 SMS tokenization problems

Tokenization process for "standard" alphabetic languages is defined as the division of character sequences into sentences and sentences into tokens. As tokens we consider words, numbers and every other punctuation marker. Although, Dale (2000) gives us a simple definition of text tokenization process without taking into account punctuation markers or numbers: *Tokenization is the process of breaking up the sequence of characters in a text by locating the word boundaries, the points where one word ends and another begins*.

The importance of this process for Natural Language Processing (NLP) applications such as POS taggers, parsers, search engines, text normalization etc. is because they deal with words and sentences. Most tokenizer applications use a simple method which implements words separations by blanks, thus a white space is a delimiter of word boundaries and also separate punctuation markers (Schmid, 2007). For alphabetic languages the main problem of tokenization is the ambiguity between abbreviation periods, multiword expressions, sentence markers, etc. (fig., etc., U.K., S. Africa, have fun).

It is already hard to delimit the boundaries of a "standard" alphabetic language token, with regard to SMS language we release that segmentation of tokens becomes a real "challenge". To these standard tokenization problems joins SMS tokenization problems with graphical, phonetical and morphological particularities. An SMS text is characterized by an unpredictable use of whitespaces, special characters and a lack of any writing standards. SMS word is not always surrounded by whitespaces, punctuation marks are usually absent and special marks, such as emoticons, are frequently used.

We summarize below some SMS problems which need to be solved :

- Multiword non-standard abbreviations: tokens which borrow the initials of a multiword expression ex. lol = laugh out loud, stp = s'il te plait (please)
- Sentence boundary detection: most of the time a punctuation mark is missing at the end of a SMS sentences
- Missing whitespaces and punctuation marks: abbreviations promote the omission of an apostrophe or a whitespace between two or three words which generate semantic ambiguities ex. ct= cette (this), ct= c'est (it is)
- Other punctuations – Emoticons: it's about symbolic figures composed by punctuation marks and letters which represent a graphical form of emotions ex. :) = smile, ;) = winking
- Mix of characters and numbers: SMS words are usually composed by numbers and characters ex. 2day= today, dem1= demain (tomorrow)
- Extending punctuation marks: commonly used in order to express a large wonder, admiration, the thought or happiness and sadness with emoticons ex. quoi??????? (what???????), :)))))))))

3.1 From tokenization approaches to SMS normalization

The fundamental step of a text pre-processing is the normalization of a text. Sproat el al.(2001) insist in the fact that normalization must be applied before any other classic NLP process. Most of the time, normalization involves tokenization process. As it concerns SMS, text tokenization is a trivial processing stage. Normalization process of SMS aims to convert informal text in a grammatically correct text. Non standardized SMS

message is represented as a sequence $T = T_1, T_2, ..., T_n$ of tokens. As a given token T_i, we define the operation of normalization R, such as $R(T) = r_1, r_2, ..., r_n$ is a set of normalizations of T:

Given T_i= *combien (how many)*
R (combien) = cmbien, cb, cmb, kmbien, cbien

There are three approaches till now in order to achieve an SMS normalisation : a) spell checking, b) machine translation and c) automatic speech recognition (Kobus et al., 2008). Beaufort et al. (2010) propose a hybrid rule which combines both of these approaches spell checking and machine translation. These methods are based on models learned from a SMS aligned at character level corpus and its transcription. With the purpose of tokenizing Twitter messages which are similar to SMS messages, Kaufmann and Kalita (2010) use a two step model that fist preprocess messages to remove noise and they feed them into a machine translation model in order to convert them into standard English. Although, neither Kobus et al.(2008) nor Kaufmann and Kalita (2010) take into account phonetic similarities which are frequently presented. Han et al. (2011), at the other side, use a cascaded method which detects bad-formed words and generates candidates based on morphophonemic similarities. An alternative approach offers Aw et al. (2006), by a different point of view, he consider normalization as a translation problem and adopt a method which aims to adapt a phrase based statistical machine translation model. Choudhury et al. (2007) propose the application of a model in which the system of normalization uses statistical methods spelling correction conversion based on HMM (Hidden Markov Models) between texting and the standard language. This model was used to construct a decoder SMS text in English to their standard English forms with an accuracy of 89% at the word level. On the same model is based Lopez et al. (2014) in order to obtain a semi-automatic alignment method messages in order to build a dictionary SMS.

Most of the applied studies are based on deterministic techniques for automatic construction of transcription dictionaries, statistical methods for the automatic transcription of a SMS word and analysis of hybrid approaches (deterministic-probabilistic). Our aim is to focus on transcription process from SMS messages to standard french language. As starting point, of our research we consider that every SMS word refers to a standard language word and there is always a standard word definition for SMS words. We examine multiple different graphical forms of a SMS word by giving the definition of the term *polygraphy* which means that a SMS word can be transcribed in two or more standard words. At the same time, a standard french word can be transcribed in two or more SMS words. Of course, we couldn't omit the fact of the correspondence of one SMS word to one standard word. To this day, these graphical aspects are poorly developed in the SMS related literature (Fairon and Paumier, 2006; Beaufort et al., 2010; Cougnon and François, 2011; Panckhurst, 2009). These observations permit us to have a global view of the ambiguity level that we face in SMS transcription. The goal of our study is to achieve a transcription approach of SMS words to standard language word by applying a rule-based model.

4 Conclusion

In this paper we have presented the alpes4science project from the collection to the processing of SMS messages. Based on SMS language particularities we had defined the tokenization problems and penetrate into normalizations approaches. The alpes4science database is a composition of 22,054 authentic text messages which had been semi-automatically proceed. As a result we dispose an aligned corpus of SMS messages with their transcription, anonymization and segmentation, a dictionary with the couple of SMS words and translation and metadata of the participants' social profile. This material composes an indisputable tool for sociolinguistic and linguistic researches, as well as for NLP applications (automatic name entity extraction, normalization, information retrieval etc.). The processing of the SMS corpus allows us this day to expect the upcoming online publication of the corpus by the Consortium of written corpus, of CoMeRe project.

5 Acknowledgment

Funding for this project was provided by a grant from *la Région Rhône-Alpes*.

References

Antoniadis G., Chabert G., and Zampa V. 2011. Alpes4science: Constitution d'un corpus de SMS réels en France métropolitaine, talk. *79th Acfas colloquium*, Sherbrooke, May 9-10, 2011.

Aw A., Zhang M., Xiao J. and Su J. 2006. A phrase-based statistical model for SMS text normalization. In *Proc. COLING/ACL 2006*.

Barasa S. and Mous M. 2009. The Oral & Written Interface in SMS: Technologically Mediated Communication in Kenya. *Low Educated Second Language and Literacy Acquisition*.

Beaufort R., Roekhaut S., Cougnon L.-A., Fairon C. 2010. Une approche hybride traduction/correction pour la normalisation des SMS Richard. In *TALN 2010*.

Chabert G., Zampa V., Antoniadis G., Mallen, M. 2012. *Des SMS Alpins* Editions de la Bibliotheque départementale de Hautes-Alpes.

Choudhury M., Saraf R., Jain V., Mukherjee A., Sarkar1 S., and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157– 174.

Cougnon, L.-A., and François, T. 2011. Etudier l'écrit SMS - Un objectif du projet sms4science. In *Linguistik online 48*.

Dale, R. 2000. *Handbook of Natural Language Processing* (p. 964).

Fairon C. and Paumier S. 2006. *Le langage SMS* . Louvain-la-Neuve: Presses universitaires de Louvain.

Han, B., and Baldwin, T. 2011. Lexical Normalisation of Short Text Messages : Makn Sens a # twitter. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* .

Kaufmann J. and Kalita J. 2010. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India.

Kobus, C., Marzin, P., and Lannion, F. 2008. Normalizing SMS : are two metaphors better than one ? *In: COLING 2008*.

Lopez C., Bestandji R., Roche M. and Panckhurst R. 2014. Towards Electronic SMS Dictionary Construction: An Alignment-based Approach *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Panckhurst R. 2013. A Large SMS Corpus in French: From Design and Collation to Anonymisation, Transcoding and Analysis. In *5th International Conference on Corpus Linguistics (CILC2013)*.

Panckhurst R. 2009. Short Message Service (SMS) : typologie et problématiques futures, in *Arnavielle T. (coord.), 33–52*.

Sproat R., Black A.W., Chen S., Kumar S. , Ostendorf M. , and Richards C. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Stark E. 2011. La morphosyntaxe dans les SMS suisses francophones: Le marquage de l'accord sujet – verbe conjugué *Linguistik Online, 48(4):35-47*.

Schmid, H. 2007. Tokenizing. *Anke Lüdeling and Merja Kytö* (Corpus Lin., pp. 1–17). Mouton de Gruyter, Berlin.